Protein Binding Affinity Prediction Using Support Vector Regression and Interfecial Features

Adiba Yaseen
Department of Computer and information Science
Pakistan Institute of Engineering and applied Science (PIEAS)
Islamabad, pakistan
adibayaseen@gmail.com

Wajid Arshad Abbasi
Department of Computer and information Science
Pakistan Institute of Engineering and applied Science (PIEAS)
Islamabad, pakistan
wajidarshad@gmail.com

Fayyaz ul Amir Afsar Minhas
Department of Computer and information Science
Pakistan Institute of Engineering and applied Science (PIEAS)
Islamabad, pakistan
afsar@pieas.edu.pk

***Abstract*— In understanding biology at the molecular level, analysis of protein interactions and protein binding affinity is a challenge. It is an important problem in computational and structural biology. Experimental measurement of binding affinity in the wet-lab is expensive and time consuming. Therefore, machine learning approaches are widely used to predict protein interactions and binding affinities by learning from specific properties of existing complexes. In this work, we propose an innovative computational model to predict binding affinities and interaction based on sequence, structural and interface features of the interacting proteins that are robust to binding associated conformational changes.**

We modeled the prediction of binding affinity as classification and regression problem with least-squared and support vector regression models using structure and sequence features of proteins. Specifically, we have used the number and composition of interacting residues at protein complexes interface as features and sequence features. We evaluated the performance of our prediction models using Affinity Benchmark Dataset version 2.0 which contains a diverse set of both bound and unbound protein complex structures with known binding affinities. We evaluated our regression performance results with root mean square error (RMSE) as well as Spearman and Pearson's correlation coefficients using a leave-one-out cross-validation protocol. We evaluate classification results with* AUC-ROC *and AUC-PR. Our results show that Support Vector Regression performs significantly better than other models with a Spearman Correlation coefficient of 0.58, Pearson Correlation score of 0.55 and RMSE of 2.41 using 3-mer and sequence feature. It is interesting to note that simple features based on 3-mer features and the properties of the interface of a protein complex are predictive of its binding affinity. These features, together with support vector regression achieve higher accuracy than existing sequence based methods.

Keywords—Support vector machine (SVM), Area under the Receiver Operating Characteristic Curve (AUC-ROC), Area under the Precision-Recall Curve (AUC-PR)

I. INTRODUCTION

Proteins are long linear chains of amino acids that form the basis of life [1]. Protein molecules do not function in isolation and most of their functions are executed through interactions or binding with other proteins [2].When two proteins bind each other, the strength of their binding is called binding affinity. Binding affinity is an important parameter in the study of proteins, especially in protein interaction prediction, drug design, computation protein design, etc. [3]. It also plays a major role in understanding the relationship between structure and function of protein complexes and biochemical pathways[4].

There are various experimental and biophysical methods for measurement of binding affinity like Nuclear magnetic resonance spectroscopy, Surface Plasmon Resonance (SPR), pull-down assays, analytical ultracentrifugation and some other methods [5]. However, these experimental methods are not applicable on a large-scale due to high costs in terms of time, money and personnel. An accurate computational model for binding affinity prediction can solves this issue. As a consequence, binding affinity prediction from sequence and structure of proteins is an important research problem.

Computation methods for binding affinity prediction can be divided into three categories: Force field or molecular simulation based methods, docking and machine learning methods [6]–[8]. Due to their high computational cost, existing scoring function methods are trained and tested using small datasets and are not applicable on a large scale. Docking methods are applicable only if proteins do not undergo large changes in their structural conformations from unbound to bound state. Among computational methods for binding affinity prediction, machine learning methods are particularly suitable due to flexible learning function and better generalization performance [9]

Machine learning models learn from existing properties of data and predict affinity values for unseen data. Every machine learning model needs diverse dataset for training, testing and cross validation. After preprocessing of dataset features are extracted and a regression or classification model is trained to be used for unseen data. A number of machine learning based binding affinity prediction methods use the binding affinity benchmark data[10]. Most existing methods work with protein structures or features derived from protein complex interfaces[8], [11]–[13]. However, these methods are limited by the requirement that high resolution protein structures of the proteins forming the complex be available for generating a prediction. In contrast, sequence based prediction

methods are more widely applicable. However, development of effective prediction models using sequence information alone is difficult because protein binding affinity and interaction depend upon structure and function of proteins. In existing literature, the highest accuracy for sequence based binding affinity prediction has been reported for a machine learning method by Yugandar and Gromiha with a Pearson correlation score of 0.739 to 0.992 for different types of complexes [14]. However, the accuracy of their method could not be independently verified [15]. It was also found that their proposed method does not work well on an independent test set. In this paper, we build on these findings and present a machine learning model using sequence and interface features. The rest of the paper is organized as follows: Section II presents materials and methods with complete details of the proposed algorithm, section. Results and conclusions are presented in sections III and IV, respectively.

II. Materials and Methods

A. Dataset and pre-processing

We evaluated the performance of our prediction models using Affinity Benchmark Dataset (version 2.0) which contains 135 non-redundant and diverse structures of ligand and receptor proteins in both bound and unbound forms with known binding affinities. The dataset has binding affinities for all these complexes in the form of free energy ΔG (ranges from -18.58 to -4.29) and disassociation constant (K_d) [14].

B. Feature extraction

We model the prediction of binding affinity as both regression and classification problems. Specifically, least-squared and support vector regression models are used for regression using structure and sequence features of proteins. These machine learning models require meaningful features to be extracted from a given complex to predict binding affinity. We begin the description of our method by introducing various feature extraction schemes used in this work. Following features are extracted from both ligand and receptor proteins in a given protein complex and the two individual feature representations are concatenated to form a feature vector for the protein complex.

1) Amino acid composition

Amino acid composition of a protein or protein complex is a 20-dimensional vector $\varphi_{AAC}(\boldsymbol{x})$ which contains the frequency of occurrence of all 20 natural amino acids in the sequences.

2) BLOSUM-62 Average

In order to capture physiochemical similarity of amino acids forming a protein or a complex, we used the average of the rows of the BLOSUM-62 substitution matrix corresponding to different amino acids in the sequence. This results in a 20-dimensional feature vector $\varphi_{BLOSUM}(\boldsymbol{x})$ for a given sequence [16].

3) 3-mer Features

In order to capture sequence characteristics, we used the frequency of occurrence of all possible 3-mers of amino acids in a given sequence. This results in an 8000-dimensional feature vector and has been useful in our previous work[17].

4) Propy features

We also extracted Propy features for a given sequence using the Propy package in Python [18]. It extracts 13 different types of features from a given protein sequence such as dipeptide composition, tripeptide composition, pseudo amino acid (PseAAC) composition, Moran autocorrelation, normalized Moreau-Broto autocorrelation, quasi-sequence-order descriptors, composition [19]. These features capture both local and global level features in a protein sequence.

5) Interfacial amino acid composition

Since protein-protein binding is a consequence of non-covalent interactions between interfacial residues of ligand and receptor proteins, features extracted from protein interface can be very useful in predicting binding affinity. For this purpose, we calculated frequency of non-repeating pairs of residues at the interface of a protein complex. It is important to note that, unlike previously discussed features, these features are extracted from the structure of a protein complex.

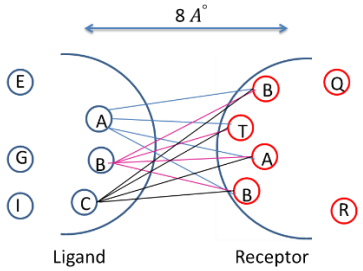

Figure 1 Number of Interacting Interface Features that lie within 8 Angstorm distance between ligand and receptor residues.

Specifically, a 211-dimensional feature vector is formed with each element corresponding to the number of times a specific amino acid pair occurs in the interface. Interface residues were identified as residues that occur on the two proteins in a protein complex within a distance of 8 Angstroms from each other as shown in the Figure 1.

C. Classification

Initially we modelled binding affinity prediction as a classification problem in which the objective was to classify a given protein complex into low (below median) or high (above median) affinity classes. For this purpose, we used a support vector machine classifier (SVM) with a radial basis function kernel. Given a training dataset containing N protein complexes each with a feature representation $\boldsymbol{\varphi}(x_i)$ and known class label $y_i \in \{-1, +1\}$, a support vector machine learns a discriminant function $f(\boldsymbol{x}) = \boldsymbol{w}^T \boldsymbol{\varphi}(x) + b$ by solving the following optimization problem:

$$\min_{\boldsymbol{w},b,\xi \geq 0} \frac{1}{2}||\boldsymbol{w}||^2 + C\sum_{i=1}^{N} \xi_i$$

$$\text{Such that for all } i\text{: } y_i(\boldsymbol{w}^T \boldsymbol{\varphi}(x_i) + b) \geq 1 - \xi_i$$

In this optimization problem, the objective is to find the weight vector $\boldsymbol{w}$ and bias b for the decision function such that the classification constraints are satisfied for the training data.

The hyper-parameter C controls the extent of regularization vs. training error minimization and is selected through cross-validation. An SVM can also perform non-linear classification through the use of kernel functions such as a radial basis function [20].

D. Regression Methods

We also modelled the binding affinity prediction problem using linear and support vector regression as discussed below. In these formulations, the real value of binding affinity is used as the target y_i for a complex instead of binary (low or high) labels as in the classification problem.

1) Linear regression

Linear regression fits least squared linear function to a given data set by solving the following optimization problem: $min_{w,b} \sum_{i=1}^{N} (\boldsymbol{w^T \varphi}(x_i) + b - y_i)^2$. We used least squared regression as a baseline due to its limitations such as sensitivity to outliers, inability to account for non-linear relationships and degradation in performance in high dimensions. Given the feature representation of a test protein complex $\boldsymbol{\varphi}(x)$, its predicted binding affinity is calculated by the function: $f(\boldsymbol{x}) = \boldsymbol{w}^T \boldsymbol{\varphi}(x) + b$.

2) Support vector regression

In order to overcome the limitations of simple linear regression, we have also used support vector regression (SVR). Due to the use ϵ-insestive loss instead of least-squared error, SVR is more robust to outliers and can perform non-linear regression using kernel functions [21]. Specifically, SVR works by solving the following optimization problem:

$$\min_{\boldsymbol{w},b,\boldsymbol{\xi} \geq 0} \frac{1}{2} ||\boldsymbol{w}||^2 + C \sum_{i=1}^{N} \xi_i$$

Such that for all i: $|(\boldsymbol{w}^T \boldsymbol{\varphi}(x_i) + b) - y_i| \leq \epsilon + \xi_i$

In this formulation, the first term is the regularization term and it controls the capacity of the machine learning model whereas the second term performs error minimization. Similar to an SVM classifier, the hyper-parameter C controls the extent of regularization.

E. Performance Evaluation and Hyper-parameter selection

The proposed models were implemented in Python 2.7. We have used leave-one-out cross validation (LOOCV) for evaluating our machine learning models. In LOOCV, a machine learning model is trained on all but one of the complexes which is used for testing and this procedure is repeated for all complexes and then the performance metrics are evaluated by comparing the predictions to known labels or target values. Hyper-parameters (such as C or the spread parameter in radial basis function kernels) are obtained by nested grid search. In order to evaluate the performance of our classification model, we used the following performance metrics:

- *Area Under the Receiver Operating Characteristic Curve (AUC-ROC)*: This metric captures the quality of classification by calculating the area under the curve formed by plotting the true positive rate vs. the false positive rate. A perfect classifier will have an AUC-ROC score of 1.0 whereas a random classifier will have AUC-ROC of 0.5.
- *Area Under the Precision-Recall Curve (AUC-PR)*: AUC-PR is obtained by calculating the area under the curve obtained by plotting precision of the classifier vs. recall or true positive rate. A perfect classifier will have AUC-PR of 1.0.

The performance of regression models is evaluated using the following metrics:

- *Root Mean Square Error (RMSE)*: RMSE calculates the average error between the true and predicted binding affinities of test complexes as follows:

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{N} (f(x_i) - y_i)^2}{N}}$$

- *Pearson Correlation Coefficient*: Pearson correlation coefficient captures the degree of correspondence between the true and predicted binding affinity values. A value close to 1.0 indicates very good correspondence or linear dependence between the two entities whereas a value of 0.0 indicates no correlation between them.
- *Spearman Correlation Coefficient*: Spearman correlation calculates the degree of linear dependence between the ranks of true and predicted binding affinity values of complexes. Similar to Pearson correlation coefficient, the Spearman correlation coefficient also indicates very good correspondence between true and predicted binding affinities if its value is close to 1.0.

III. RESULTS AND DISCUSSION

We evaluate the performance of three types of models, SVM, SVR and linear regression using one interface and four sequence features. Table 1 presents (LOOCV) SVM classification results using sequence features amino acid composition (AAC), 3-mer, Propy, BLOSUM and interface features together with various feature combinations. It is interesting to note that the highest accuracy is obtained with 3-mer features (AUC-ROC: 0.69, AUC-PR: 0.93) which is higher than interface features (AUC-ROC: 0.67, AUC-PR: 0.92). Among combinations of features, amino acid composition and interface features given the highest accuracy. However, no significant improvement in terms of AUC-PR is recorded.

We evaluate regression model using RMSE and Spearman correlation S_r and Pearson correlation P_r for linear regression and SVR. It is clear from Table 2 that linear regression is unable to predict binding affinity effectively. This is because linear regression is affected by outliers and produces larger errors when the number of feature dimensions is large. SVR results shows that single interface features are most informative as expected with RMSE: 2.47, S_r: 0. 55 and P_r: 0 .52. Although performance of sequence features are slightly less than interface feature but these features are

considered more important because they do not require protein complex structures to be available. 3-mer sequence feature performs better than all other sequence feature with RMSE: 2.55, S_r: 0.50 and P_r: 0.45.

Results in the Table-2 also demonstrate that combination of sequence and interfacial features (interface and 3-mer) performs better than sequence features alone with RMSE: 2.41, S_r: 0.58 and P_r: 0.55. Figure 2 shows these results in the form of a scatter plot of true vs. predicted binding affinity values. Results of both classification and regression models show that our model performs better using only sequence features than the previous state of the art [12] as reported by [15] which gives a Pearson correlation coefficient of 0.07.

IV. CONCLUSIONS & FUTURE WORK

In this work, we have presented a machine learning based methods for protein binding affinity prediction. We have shown that simple sequence based features can be used to predict binding affinity with only a minimal loss in performance in comparison to using interfacial amino acid composition which requires availability of structures of protein complexes. The proposed model can be used for prediction of binding affinities of novel protein complexes and can be improved by extracting evolutionary features. We aim to develop a webserver for the proposed model in the future. The code for the proposed model will also be made publicly available upon acceptance of this paper.

Table 1 Leave one out Cross validation SVM Results

Features	SVM	
	AUC-ROC	***AUC-PR***
AAC	0.62	0.91
BLOSUM	0.64	0.91
3-mer	**0.69**	**0.93**
Propy	**0.69**	**0.92**
Interface	0.67	0.92
Interface + 3-mer	0.72	0.92
Interface + Propy	0.72	0.92
Interface + AAC	**0.75**	**0.93**
Interface + BLOSUM	0.72	0.92
3-mer + Propy	0.70	0.92
3-mer +AAC	0.66	0.92
3-mer + BLOSUM	0.67	0.91
Propy + BLOSUM	0.66	0.91
Propy + AAC	0.66	0.91
BLOSUM + AAC	0.71	0.91

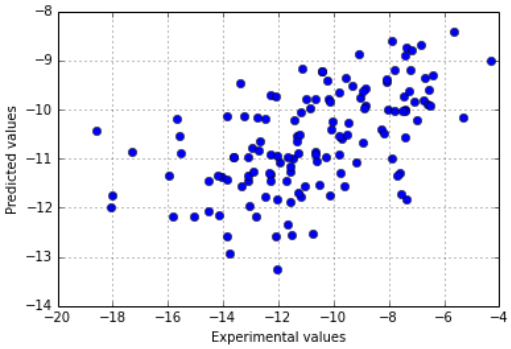

Figure 2 Correlation between Experimental and Predicted Binding Affinity values using Interface and 3-mer feature of sequence. Spearman correlation of 0.58, Pearson correlation of 0.55 and RMSE of 2.41 were recorded.

Table 2 Leave one out Cross validation Regression Results

Features	***SVR***			***Linear Regression***		
	RMSE	S_r	P_r	***RMSE***	S_r	P_r
3-mer	**2.55**	**0.50**	**0.45**	**-**	**0.18**	**0.11**
Propy	2.56	0.45	0.44	-	-0.11	-0.03
BLOSUM	2.60	0.41	0.39	3.15	0.21	0.21
AAC	2.61	0.39	0.39	3.00	0.25	0.27
Interface	**2.47**	**0.55**	**0.52**	**-**	**0.20**	**0.09**
Interface + 3-mer	**2.41**	**0.58**	**0.55**	**-**	**0.002**	**0.02**
Interface+ Propy	2.41	0.57	0.54	-	-0.05	-0.10
Interface+ AAC	2.42	0.55	0.53	4.42	0.22	0.23
Interface+ BLOSUM-62	2.49	0.51	0.47	4.43	0.18	0.18
3-mer+ Propy	2.58	0.49	0.45	-	-0.03	0.01
3-mer+AAC	2.56	0.49	0.43	3.23	0.31	0.29
3-mer+ BLOSUM	2.58	0.44	0.41	3.34	0.28	0.25

REFERENCES

[1] J. R. Perkins, I. Diboun, B. H. Dessailly, J. G. Lees, and C. Orengo, "Transient Protein-Protein Interactions: Structural, Functional, and Network Properties," *Structure*, vol. 18, no. 10, pp. 1233–1243, Oct. 2010.

[2] S. A. Hassan, L. Gracia, G. Vasudevan, and P. J. Steinbach, "Computer simulation of protein-ligand interactions: challenges and applications," *Methods Mol. Biol. Clifton NJ*, vol. 305, pp. 451–492, 2005.

[3] H. Jubb, T. L. Blundell, and D. B. Ascher, "Flexibility and small pockets at protein–protein interfaces: New insights into druggability," *Prog. Biophys. Mol. Biol.*, vol. 119, no. 1, pp. 2–9, Oct. 2015.

[4] B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter, *Molecular Biology of the Cell*, 4th ed. Garland Science, 2002.

[5] G. Sudha, R. Nussinov, and N. Srinivasan, "An overview of recent advances in structural bioinformatics of protein-protein interactions and a guide to their principles," *Prog. Biophys. Mol. Biol.*, vol. 116, no. 2–3, pp. 141–150, Dec. 2014.

[6] G. L. Warren *et al.*, "A critical assessment of docking programs and scoring functions," *J. Med. Chem.*, vol. 49, no. 20, pp. 5912–5931, Oct. 2006.

[7] J. Qin, M. J. Li, P. Wang, M. Q. Zhang, and J. Wang, "ChIP-Array: combinatory analysis of ChIP-seq/chip and microarray gene expression data to discover direct/indirect targets of a transcription factor," *Nucleic Acids Res.*, vol. 39, no. Web Server issue, pp. W430-436, Jul. 2011.

[8] A. Vangone and A. M. Bonvin, "Contacts-based prediction of binding affinity in protein–protein complexes," *eLife*, vol. 4, p. e07454, Jul. 2015.

[9] Q. U. Ain, A. Aleksandrova, F. D. Roessler, and P. J. Ballester, "Machine-learning scoring functions to improve structure-based binding affinity prediction and virtual screening," *Wiley Interdiscip. Rev. Comput. Mol. Sci.*, vol. 5, no. 6, pp. 405–424, Dec. 2015.

[10] P. L. Kastritis *et al.*, "A structure-based benchmark for protein-protein binding affinity," *Protein Sci. Publ. Protein Soc.*, vol. 20, no. 3, pp. 482–491, Mar. 2011.

[11] F. Tian, Y. Lv, and L. Yang, "Structure-based prediction of protein–protein binding affinity with consideration of allosteric effect," *Amino Acids*, vol. 43, no. 2, pp. 531–543, Aug. 2012.

[12] I. H. Moal, R. Agius, and P. A. Bates, "Protein–protein binding affinity prediction on a diverse set of structures," *Bioinformatics*, vol. 27, no. 21, pp. 3002–3009, Nov. 2011.

[13] A. Erijman, E. Rosenthal, and J. M. Shifman, "How Structure Defines Affinity in Protein-Protein Interactions," *PLoS ONE*, vol. 9, no. 10, Oct. 2014.

[14] K. Yugandhar and M. M. Gromiha, "Protein–protein binding affinity prediction from amino acid sequence," *Bioinformatics*, vol. 30, no. 24, pp. 3583–3589, Dec. 2014.

[15] K. Yugandhar and M. M. Gromiha, "Response to the comment on 'protein-protein binding affinity prediction from amino acid sequence,'" *Bioinformatics*, vol. 31, no. 6, pp. 978–978, Mar. 2015.

[16] S. R. Eddy, "Where did the BLOSUM62 alignment score matrix come from?," *Nat. Biotechnol.*, vol. 22, no. 8, pp. 1035–1036, Aug. 2004.

[17] T. Zhou *et al.*, "Quantitative modeling of transcription factor binding specificities using DNA shape," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 112, no. 15, pp. 4654–4659, Apr. 2015.

[18] D.-S. Cao, Q.-S. Xu, and Y.-Z. Liang, "propy: a tool to generate various modes of Chou's PseAAC," *Bioinformatics*, vol. 29, no. 7, pp. 960–962, Apr. 2013.

[19] I. Limongelli, S. Marini, and R. Bellazzi, "PaPI: pseudo amino acid composition to score human protein-coding variants," *BMC Bioinformatics*, vol. 16, p. 123, Apr. 2015.

[20] "An overview of statistical learning theory - IEEE Journals & Magazine." [Online]. Available: http://ieeexplore.ieee.org/abstract/document/788640/. [Accessed: 01-Oct-2017].

[21] A. J. Smola and B. Schölkopf, "A tutorial on support vector regression," *Stat. Comput.*, vol. 14, no. 3, pp. 199–222, Aug. 2004.